Incremental Bayesian Learning of Semantic Categories

Incremental Bayesian Learning of Semantic Categories

Lea Frermann and **Mirella Lapata**
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB
l.frermann@ed.ac.uk, mlap@inf.ed.ac.uk

Abstract

Models of category learning have been extensively studied in cognitive science and primarily tested on perceptual abstractions or artificial stimuli. In this paper we focus on categories acquired from *natural language* stimuli, that is words (e.g., *chair* is a member of the FURNITURE category). We present a Bayesian model which, unlike previous work, learns both categories and their features in a single process. Our model employs particle filters, a sequential Monte Carlo method commonly used for approximate probabilistic inference in an incremental setting. Comparison against a state-of-the-art graph-based approach reveals that our model learns qualitatively better categories and demonstrates cognitive plausibility during learning.

1 Introduction

Considerable psychological research has shown that people reason about novel objects they encounter by identifying the category to which these objects belong and extrapolating from their past experiences with other members of that category (Smith and Medin, 1981). *Categorization* is a classic problem in cognitive science, underlying a variety of common mental tasks including perception, learning, and the use of language.

Given its fundamental nature, categorization has been extensively studied both experimentally and in simulations. Indeed, numerous models exist as to how humans categorize objects ranging from strict *prototypes* (categories are represented by a single idealized member which embodies their core properties; e.g., Reed 1972) to full exemplar models (categories are represented by a list of previously encountered members; e.g., Nosofsky 1988) and combinations of the two (e.g., Griffiths et al. 2007). A common feature across different studies is the use of stimuli involving real-world objects (e.g., children's toys; Starkey 1981), perceptual abstractions (e.g., photographs of animals; Quinn and Eimas 1996), or artificial ones (e.g., binary strings, dot patterns or geometric shapes; Medin and Schaffer 1978; Posner and Keele 1968; Bomba and Siqueland 1983). Most existing models focus on adult categorization, in which it is assumed that a large number of categories have already been learnt (but see Anderson 1991 and Griffiths et al. 2007 for exceptions).

In this work we focus on categories acquired from *natural language* stimuli (i.e., words) and investigate how the statistics of the linguistic environment (as approximated by large corpora) influence category formation (e.g., *chair* and *table* are FURNITURE whereas *peach* and *apple* are FRUIT[1]). The idea of modeling categories using words as a stand-in for their referents has been previously used to explore categorization-related phenomena such as semantic priming (Cree et al., 1999) and typicality rating (Voorspoels et al., 2008), to evaluate prototype and exemplar models (Storms et al., 2000), and to simulate early language category acquisition (Fountain and Lapata, 2011). The idea of using naturalistic corpora has received little attention. Most existing studies use feature norms as a proxy for people's representation of semantic concepts. In a typical procedure, participants are presented with a word and asked to generate the most relevant features or attributes for its referent concept. The most notable collection of feature norms is probably the multi-year project of McRae et al. (2005), which obtained features for a set of 541 common English nouns.

Our approach replaces feature norms with representations derived from words' contexts in corpora. While this is an impoverished view of how categories are acquired — it is clear that they are learnt through exposure to the linguistic environment *and* the physical world — perceptual infor-

[1] Throughout this paper we will use small caps to denote CATEGORIES and italics for their *members*.

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 249–258,
Gothenburg, Sweden, April 26-30 2014.

mation relevant for extracting semantic categories is to a large extent redundantly encoded in linguistic experience (Riordan and Jones, 2011). Besides, there are known difficulties with feature norms such as the small number of words for which these can be obtained, the quality of the attributes, and variability in the way people generate them (see Zeigenfuse and Lee 2010 for details). Focusing on natural language categories allows us to build categorization models with theoretically unlimited scope.

To this end, we present a probabilistic Bayesian model of category acquisition based on the key idea that learners can adaptively form category representations that capture the structure expressed in the observed data. We model category induction as two interrelated sub-problems: (a) the acquisition of features that discriminate among categories, and (b) the grouping of concepts into categories based on those features. An important modeling question concerns the exact mechanism with which categories are learned. To maintain cognitive plausibility, we develop an *incremental* learning algorithm. Incrementality is a central aspect of human learning which takes place sequentially and over time. Humans are capable of dealing with a situation even if only partial information is available. They adaptively learn as new information is presented and locally update their internal knowledge state without systematically revising everything known about the situation at hand. Memory and processing limitations also explain why humans must learn incrementally. It is not possible to store and have easy access to all the information one has been exposed to. It seems likely that people store the most prominent facts and generalizations, which they modify on they fly when new facts become available.

Our model learns categories using a particle filter, a Markov Chain Monte Carlo (MCMC) inference mechanism which sequentially integrates newly observed data and can be thus viewed as a plausible proxy for human learning. Experimental results show that the incremental learner obtains meaningful categories which outperform the state of the art whilst at the same time acquiring semantic representations of words and their features.

2 Related Work

The problem of category induction has achieved much attention in the cognitive science literature. Incremental category learning was pioneered by Anderson (1991) who develops a non-parametric model able to induce categories from abstract stimuli represented by binary features. Sanborn et al. (2006) present a fully Bayesian adaptation of Anderson's original model, which yields a better fit with behavioral data. A separate line of work examines the cognitive characteristics of category acquisition as well as the processes of generalizing and generating new categories and exemplars (Jern and Kemp, 2013; Kemp et al., 2012). The above models are conceptually similar to ours. However, they were developed with adult categorization in mind, and use rather simplistic categories representing toy-domains. It is therefore not clear whether they generalize to arbitrary stimuli and data sizes. We aim to show that it is possible to acquire natural language categories on a larger scale purely from linguistic context.

Our model is loosely related to Bayesian models of word sense induction (Brody and Lapata, 2009; Yao and Durme, 2011). We also assume that local linguistic context can provide important cues for word meaning and by extension category membership. However, the above models focus on performance optimization and learn in an ideal batch mode, while incorporating various kinds of additional features such as part of speech tags or dependencies. In contrast, we develop a cognitively plausible (early) language learning model and show that categories can be acquired purely from context, as well as in an incremental fashion.

From a modeling perspective, we learn categories incrementally using a particle filtering algorithm (Doucet et al., 2001). Particle filters are a family of sequential Monte Carlo algorithms which update the state space of a probabilistic model with newly encountered information. They have been successfully applied to natural language acquisition tasks such as word segmentation (Borschinger and Johnson, 2011), or sentence processing (Levy et al., 2009). Sanborn et al. (2006) also use particle filters for small-scale categorization experiments with artificial stimuli. To the best of our knowledge, we present the first particle filtering algorithm for large-scale category acquisition from natural text.

Our work is closest to Fountain and Lapata (2011) who also develop a model for inducing natural language categories. Specifically, they propose an incremental version of Chinese Whispers (Biemann, 2006), a randomized graph-clustering algorithm. The latter takes as input a graph which is constructed from corpus-based co-occurrence statistics and produces a hard clustering over the nodes in the graph. Contrary to our model, they treat the tasks of inferring a semantic representa-

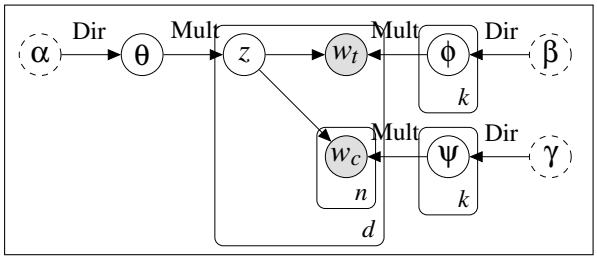


Figure 1: Plate diagram representation of the BayesCat model.

Draw distribution over categories $\theta \sim Dir(\alpha)$
for category k **do**
  Draw target word distribution $\phi_k \sim Dir(\beta)$
  Draw context word distribution $\psi_k \sim Dir(\gamma)$
for Document d **do**
  Draw category $z^d \sim Mult(\theta)$
  Draw target word $w_t^d \sim Mult(\phi_{z^d})$
  for context position $n = \{1..N\}$ **do**
    Draw context word $w_c^{d,n} \sim Mult(\psi_{z^d})$

Figure 2: The generative process of the BayesCat model.

tion for concepts and their class membership as two separate processes. This allows to experiment with different ways of initializing the co-occurrence matrix (e.g., from bags of words or a dependency parsed corpus), however at the expense of cognitive plausibility. It is unlikely that humans have two entirely separate mechanisms for learning the meaning of words and their categories. We formulate a more expressive model within a probabilistic framework which captures the meaning of words, their similarity, and the predictive power of their linguistic contexts.

3 The BayesCat Model

In this section we present our Bayesian model of category induction (BayesCat for short). The input to the model is natural language text, and its final output is a set of clusters representing categories of semantic concepts found in the input data. Like many other semantic models, BayesCat is inspired by the distributional hypothesis which states that a word's meaning is predictable from its context (Harris, 1954). By extension, we also assume that contextual information can be used to characterize *general semantic categories*. Accordingly, the input to our model is a corpus of documents, each defined as a target word t centered in a fixed-length context window:

$$[c_{-n} \ldots c_{-1} \;\; t \;\; c_1 \ldots c_n] \tag{1}$$

We assume that there exists one global distribution over categories from which all documents are generated. Each document is assigned a category label, based on two types of features: the document's target word and its context words, which are modeled through *separate* category-specific distributions. We argue that it is important to distinguish between these features, since words belonging to the same category do not necessarily co-occur, but tend to occur in the same contexts. For example, the words *polar bear* and *anteater* are both members of the category ANIMAL. However, they rarely co-occur (in fact, a cursory search using Google yields only three matches for the query "*polar bear * anteater*"). Nevertheless, we would expect to observe both words in similar contexts since both animals *eat, sleep, hunt, have fur, four legs*, and so on. This distinction contrasts our category acquisition task from the classical task of topic inference.

Figure 1 presents a plate diagram of the BayesCat model; an overview of the generative process is given in Figure 2. We first draw a global category distribution θ from the Dirichlet distribution with parameter α. Next, for each category k, we draw a distribution over target words ϕ_k from a Dirichlet with parameter β *and* a distribution over context words ψ_k from a Dirichlet with parameter γ. For each document d, we draw a category z^d, then a target word, and N context words from the category-specific distributions ϕ_{z^d} and ψ_{z^d}, respectively.

4 Learning

Our goal is to infer the joint distribution of all hidden model parameters, and observable data W. Since we use conjugate prior distributions throughout the model, this joint distribution can be simplified to:

$$P(W,Z,\theta,\phi,\psi;\alpha,\beta,\gamma) \propto \frac{\prod_k \Gamma(\mathcal{N}_k + \alpha_k)}{\Gamma(\sum_k \mathcal{N}_k + \alpha_k)} \times \prod_{k=1}^{K} \frac{\prod_r \Gamma(\mathcal{N}_r^k + \beta_r)}{\Gamma(\sum_r \mathcal{N}_r^k + \beta_r)} \times \prod_{k=1}^{K} \frac{\prod_s \Gamma(\mathcal{N}_s^k + \gamma_s)}{\Gamma(\sum_s \mathcal{N}_s^k + \gamma_s)}, \tag{2}$$

where r and s iterate over the target and context word vocabulary, respectively, and the distribu-

tions θ, ϕ, and ψ are integrated out and implicitly captured by the corresponding co-occurrence counts $\mathcal{N}^*_*$. $\Gamma()$ denotes the Gamma function, a generalization of the factorial to real numbers.

Since exact inference of the parameters of the BayesCat model is intractable, we use sampling-based approximate inference. Specifically, we present two learning algorithms, namely a Gibbs sampler and a particle filter.

The Gibbs Sampler Gibbs sampling is a well-established approximate learning algorithm, based on Markov Chain Monte Carlo methods (Geman and Geman, 1984). It operates in batch-mode by repeatedly iterating through all data points (documents in our case) and assigning the currently sampled document d a category z^d conditioned on the current labelings of all other documents z^{-d}:

$$z^d \sim P(z^d | z^{-d}, W^{-d}; \alpha, \beta, \gamma), \quad (3)$$

using equation (2) but ignoring information from the currently sampled document in all co-occurrence counts.

The Gibbs sampler can be seen as an ideal learner, which can view and revise any relevant information at any time during learning. From a cognitive perspective, this setting is implausible, since a human language learner encounters training data incrementally and does not systematically revisit previous learning decisions. Particle filters are a class of incremental, or sequential, Monte Carlo methods which can be used to model aspects of the language learning process more naturally.

The Particle Filter Intuitively, a particle filter (henceforth PF) entertains a fixed set of N weighted hypotheses (particles) based on previous training examples. Figure 3 shows an overview of the particle filtering learning procedure. At first, every particle of the PF is initialized from a base distribution P_0 (Initialization). Then a single iteration over the input data $\mathbf{y}$ is performed, during which the posterior distribution of each data point y^t under all current particles is computed given information from all previously encountered data points $\mathbf{y^{t-1}}$ (Sampling/Prediction). Crucially, each update is conditioned only on the previous model state z^{t-1}, which results in a constant state space despite an increasing amount of available data. A common problem with PF algorithms is weight degeneration, i.e., one particle tends to accumulate most of the weight. To avoid this problem, at regular intervals the set of particles is resampled in order to discard particles with low probability and to ensure that the sample is representative of the state space at any time (Resampling).

for particle p **do** ▷ Initialization
  Initialize randomly or from $z^0_p \sim p_0(z)$
for observation t **do**
  for particle n **do** ▷ Sampling/Prediction
    $P_n(z^t_n | \mathbf{y}^t) \sim p(z^t_n | z^{t-1}_n, \alpha) P(y^t | z^t_n, \mathbf{y}^{t-1})$
  $\mathbf{z}^t \propto Mult(\{P_n(z^t_n)\}^N_{i=1})$ ▷ Resampling

Figure 3: The particle filtering procedure.

This general algorithm can be straightforwardly adapted to our learning problem (Griffiths et al., 2011; Fearnhead, 2004). Each observation corresponds to a document, which needs to be assigned a category. To begin with, we assign the first observed document to category 0 in all particles (Initialization). Then, we iterate once over the remaining documents. For each particle n, we compute a probability distribution over K categories based on the simplified posterior distribution as defined in equation (2) (Sampling/Prediction), with co-occurrence counts based on the information from all previously encountered documents. Thus, we obtain a distribution over $N \cdot K$ possible assignments. From this distribution we sample with replacement N new particles, assign the current document to the corresponding category (Resampling), and proceed to the next input document.

5 Experimental Setup

The goal of our experimental evaluation is to assess the quality of the inferred clusters by comparison to a gold standard and an existing graph-based model of category acquisition. In addition, we are interested in the incremental version of the model, whether it is able to learn meaningful categories and how these change over time. In the following, we give details on the corpora we used, describe how model parameters were selected, and explain our evaluation procedure.

5.1 Data

All our experiments were conducted on a lemmatized version of the British National Corpus (BNC). The corpus was further preprocessed by removing stopwords and infrequent words (occurring less than 800 times in the BNC).

The model output was evaluated against a gold standard set of categories which was created by collating the resources developed by Fountain and

Lapata (2010) and Vinson and Vigliocco (2008). Both datasets contain a classification of nouns into (possibly multiple) semantic categories produced by human participants. We therefore assume that they represent psychologically salient categories which the cognitive system is in principle capable of acquiring. After merging the two resources, and removing duplicates we obtained 42 semantic categories for 555 nouns. We split this gold standard into a development (41 categories, 492 nouns) and a test set (16 categories, 196 nouns).[2]

The input to our model consists of short chunks of text, namely a target word centered in a symmetric context window of five words (see (1)). In our experiments, the set of target words corresponds to the set of nouns in the evaluation dataset. Target word mentions and their context are extracted from the BNC.

5.2 Parameters for the BayesCat Model

We optimized the hyperparameters of the BayesCat model on the development set. For the particle filter, the optimal values are $\alpha = 0.7, \beta = 0.1, \gamma = 0.1$. We used the same values for the Gibbs Sampler since it proved insensitive to hyperparameter variations. We run the Gibbs sampler for 200 iterations[3] and report results averaged over 10 runs. For the PF, we set the number of particles to 500, and report final scores averaged over 10 runs. For evaluation, we take the clustering from the particle with the highest weight [4].

5.3 Model Comparison

Chinese Whispers We compared our approach with Fountain and Lapata (2011) who present a non-parametric graph-based model for category acquisition. Their algorithm incrementally constructs a graph from co-occurrence counts of target words and their contexts (they use a symmetric context window of five words). Target words constitute the nodes of the graph, their co-occurrences are transformed into a vector of positive PMI values, and graph edges correspond to the cosine similarity between the PMI-vectors representing any two nodes. They use Chinese Whispers (Biemann, 2006) to partition a graph into categories.

We replicated the bag-of-words model presented in Fountain and Lapata (2011) and assessed its performance on our training corpora and test sets. The scores we report are averaged over 10 runs.

Chinese Whispers can only make hard clustering decisions, whereas the BayesCat model returns a soft clustering of target nouns. In order to be able to compare the two models, we convert the soft clusters to hard clusters by assigning each target word w to category c such that $cat(w) = \max_c P(w|c) \cdot P(c|w)$.

LDA We also compared our model to a standard topic model, namely Latent Dirichlet Allocation (LDA; Blei et al. 2003). LDA assumes that a document is generated from an individual mixture over topics, and each topic is associated with one word distribution. We trained a batch version of LDA using input identical to our model and the Mallet toolkit (McCallum, 2002).

Chinese Whispers is a parameter-free algorithm and thus determines the number of clusters automatically. While the Bayesian models presented here are parametric in that an upper bound for the potential number of categories needs to be specified, the models themselves decide on the specific value of this number. We set the upper bound of categories to 100 for LDA as well as the batch and incremental version of the BayesCat model.

5.4 Evaluation Metrics

Our aim is to learn a set of clusters each of which corresponds to one gold category, i.e., it contains *all* and *only* members of that gold category. We report evaluation scores based on three metrics which measure this tradeoff. Since in unsupervised clustering the cluster IDs are meaningless, all evaluation metrics involve a mapping from induced clusters to gold categories. The first two metrics described below perform a cluster-based mapping and are thus not ideal for assessing the output of soft clustering algorithms. The third metric performs an item-based mapping and can be directly used to evaluate soft clusters.

Purity/Collocation are based on member overlap between induced clusters and gold classes (Lang and Lapata, 2011). Purity measures the degree to which each cluster contains instances that share the same gold class, while collocation measures the degree to which instances with the same gold class are assigned to a single cluster. We report the harmonic mean of purity and collocation

[2] The dataset is available from `www.frermann.de/data`.

[3] We checked for convergence on the development set.

[4] While in theory particles should be averaged, we found that eventually they became highly similar — a common problem known as *sample impoverishment*, which we plan to tackle in the future. Nevertheless, diversity among particles is present in the initial learning phase, when uncertainty is greatest, so the model still benefits from multiple hypotheses.

as a single measure of clustering quality.

V-Measure is the harmonic mean between homogeneity and collocation (Rosenberg and Hirschberg, 2007). Like purity, V-Measure performs cluster-based comparisons but is an entropy-based method. It measures the conditional entropy of a cluster given a class, and vice versa.

Cluster-F1 is an item-based evaluation metric which we propose drawing inspiration from the supervised metric presented in Agirre and Soroa (2007). Cluster-F1 maps each target word type to a gold cluster based on its soft class membership, and is thus appropriate for evaluation of soft clustering output. We first create a $K \times G$ soft mapping matrix $\mathcal{M}$ from each induced category k_i to gold classes g_j from $P(g_j|k_i)$. We then map each target word type to a gold class by multiplying its probability distribution over soft clusters with the mapping matrix $\mathcal{M}$, and taking the maximum value. Finally, we compute standard precision, recall and F1 between the mapped system categories and the gold classes.

6 Results

Our experiments are designed to answer three questions: (1) How do the induced categories fare against gold standard categories? (2) Are there performance differences between BayesCat and Chinese Whispers, given that the two models adopt distinct mechanisms for representing lexical meaning and learning semantic categories? (3) Is our incremental learning mechanism cognitively plausible? In other words, does the quality of the induced clusters improve over time and how do the learnt categories differ from the output of an ideal batch learner?

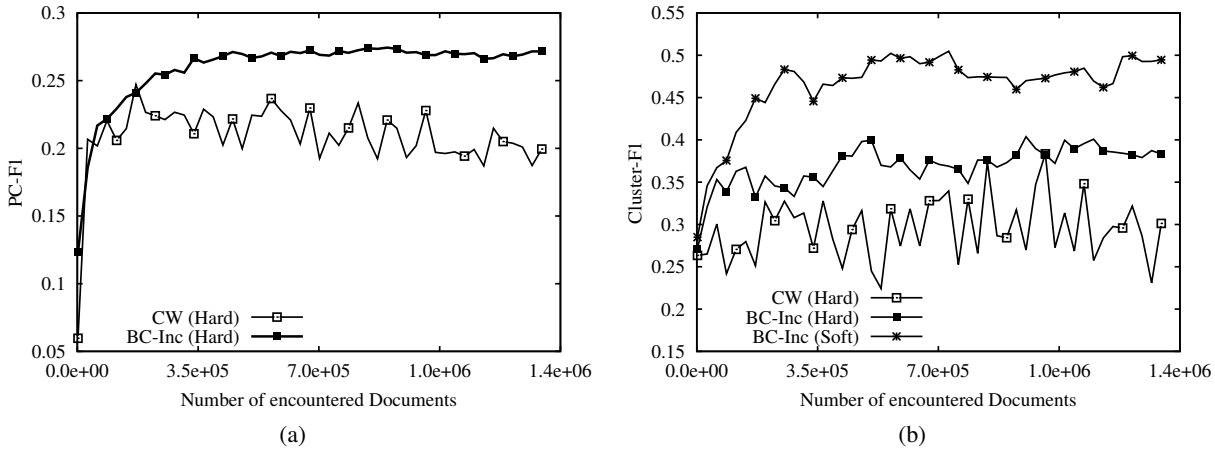
Clustering performance for the batch BayesCat model (BC-Batch), its incremental version (BC-Inc), Chinese Whispers (CW), and LDA is shown in Table 1. Comparison of the two incremental models, namely BC-Inc and CW, shows that our model outperforms CW under all evaluation metrics both on the test and the development set. Our BC models perform at least as well as LDA, despite the more complex learning objective. Recall that LDA does not learn category specific features. BC-Batch performs best overall, however this is not surprising. The BayesCat model learnt in batch mode uses a Gibbs sampler which can be viewed as an ideal learner with access to the entire training data at any time, and the ability to systematically revise previous decisions. This puts the incremental variant at a disadvantage since the particle filter encounters the data incrementally and never resamples previously seen documents. Nevertheless, as shown in Table 1 BC-Inc's performance is very close to BC-Batch. BC-Inc outperforms the Gibbs sampler in the PC-F1 metric, because it achieves higher collocation scores. Inspection of the output reveals that the Gibbs sampler induces larger clusters compared to the particle filter (as well as less distinct clusters). Although the general pattern of results is the same on the development and test sets, absolute scores for all systems are higher on the test set. This is expected, since the test set contains less categories with a smaller number of exemplars and more accurate clusterings can be thus achieved (on average) more easily.

Figure 4 displays the learning curves produced by CW and BC-Inc under the PC-F1 (left) and Cluster-F1 (right) evaluation metrics. Under PC-F1, CW produces a very steep initial learning curve which quickly flattens off, whereas no learning curve emerges for CW under Cluster-F1. The BayesCat model exhibits more discernible learning curves under both metrics. We also observe that learning curves for CW indicate much more variance during learning compared to BC-Inc, irrespectively of the evaluation metric being used. Figure 4b shows learning curves for BC-Inc when its output classes are interpreted in two ways, i.e., as soft or hard clusters. Interestingly, the two curves have a similar shape which points to the usefulness of Cluster-F1 as an evaluation metric for both types of clusters.

In order to better understand the differences in the learning process between CW and BC-Inc we tracked the evolution of clusterings over time, as well as the variance across cluster sizes at each point in time. The results are plotted in Figure 5. The top part of the figure compares the number of clusters learnt by the two models. We see that the number of clusters inferred by CW drops over time, but is closer to the number of clusters present in the gold standard. The final number of clusters inferred by CW is 26, whereas PF-Inc infers 90 clusters (there are 41 gold classes). The middle plot shows the variance in cluster size induced at any time by CW which is by orders of magnitude higher than the variance observed in the output of BayesCat (bottom plot). More importantly, the variance in BayesCat resembles the variance present in the gold standard much more closely. The clusterings learnt by CW tend to consist of

Metric		Development Set				Test Set			
		LDA	CW	BC-Inc	BC-Batch	LDA	CW	BC-Inc	BC-Batch
PC-F1	(Hard)	0.283	0.211	0.283	0.261	0.446	0.380	0.503	0.413
V-Measure	(Hard)	0.399	0.143	0.383	0.428	0.572	0.220	0.567	0.606
Cluster-F1	(Hard)	0.416	0.301	0.386	0.447	0.521	0.443	0.671	0.693
Cluster-F1	(Soft)	0.387	—	0.484	0.523	0.665	—	0.644	0.689

Table 1: Evaluation of model output against a gold standard. Results are reported for the BayesCat model trained incrementally (BC-Inc) and in batch mode (BC-Batch), and Chinese Whispers (CW). The type of clusters being evaluated is shown within parentheses.



(a) (b)

Figure 4: Learning curves for BC-Inc and CW based on PC-F1 (left), and Cluster-F1 (right). The type of clusters being evaluated is shown within parentheses. Results are reported on the development set.

few very large clusters and a large number of very small (mostly singleton) clusters. Although some of the bigger clusters are meaningful, the overall structure of clusterings does not faithfully represent the gold standard.

Finally, note that in contrast to CW and LDA, the BayesCat model learns not only how to induce clusters of target words, but also information about their category-specific contexts. Table 2 presents examples of the learnt categories together with their most likely contexts. For example, one of the categories our model discovers corresponds to BUILDINGS. Some of the context words or features relating to buildings refer to their location (e.g., *city, road, hill, north, park*), architectural style (e.g., *modern, period, estate*), and material (e.g., *stone*).

7 Discussion

In this paper we have presented a Bayesian model of category acquisition. Our model learns to group concepts into categories as well as their features (i.e., context words associated with them). Category learning is performed incrementally, using a particle filtering algorithm which is a natural choice for modeling sequential aspects of language learning.

We now return to our initial questions and summarize our findings. Firstly, we observe that our incremental model learns plausible linguistic categories when compared against the gold standard. Secondly, these categories are qualitatively better when evaluated against Chinese Whispers, a closely related graph-based incremental algorithm. Thirdly, analysis of the model's output shows that it simulates category learning in two important ways, it consistently improves over time and can additionally acquire category features.

Overall, our model has a more cognitively plausible learning mechanism compared to CW, and is more expressive, as it can simulate both category and feature learning. Although CW ultimately yields some meaningful categories, it does not acquire any knowledge pertaining to their features. This is somewhat unrealistic given that humans are good at inferring missing features for

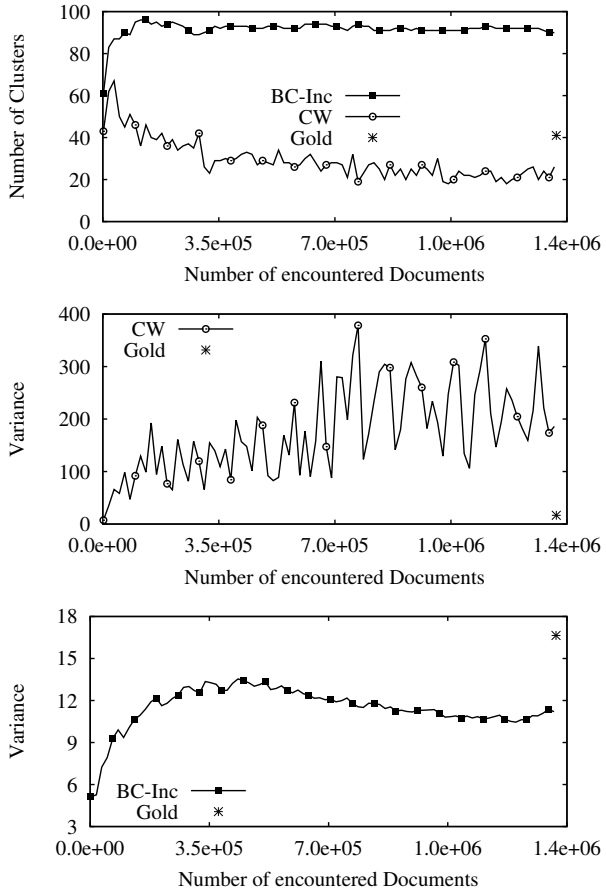


Figure 5: Number of clusters over time (top). Cluster size variance for CW (middle) and BC-Inc (bottom). Results shown on the development set.

BUILDINGS
wall, bridge, building, cottage, gate, house, train, bus, stone, chapel, brick, cathedral
plan, include, park, city, stone, building, hotel, lead, road, hill, north, modern, visit, main, period, cathedral, estate, complete, site, owner, parish

WEAPONS
shotgun, pistol, knife, crowbar, gun, sledgehammer, baton, bullet, motorcycle, van, ambulance
injure, ira, jail, yesterday, arrest, stolen, fire, officer, gun, police victim, hospital, steal, crash, murder, incident, driver, accident, hit

INSTRUMENTS
tuba, drum, harmonica, bagpipe, harp, violin, saxophone, rock, piano, banjo, guitar, flute, harpsichord, trumpet, rocker, clarinet, stereo, cello, accordion
amp, orchestra, sound, electric, string, sing, song, drum, piano, condition, album, instrument, guitar, band, bass, music

Table 2: Examples of categories induced by the incremental BayesCat model (upper row), together with their most likely context words (lower row).

unknown categories (Anderson, 1991). It is also symptomatic of the nature of the algorithm which does not have an explicit learning mechanism. Each node in the graph iteratively adopts (in random order) the strongest class in its neighborhood (i.e., the set of nodes with which it shares an edge). We also showed that LDA is less appropriate for the category learning task on account of its formulation which does not allow to simultaneously acquire clusters and their features.

There are several options for improving our model. The learning mechanism presented here is the most basic of particle methods. A common problem in particle filtering is sample impoverishment, i.e., particles become highly similar after a few iterations, and do not optimally represent the sample space. More involved resampling methods such as stratified sampling or residual resampling, have been shown to alleviate this problem (Douc, 2005).

From a cognitive perspective, the most obvious weakness of our algorithm is its strict incrementality. While our model simulates human memory restrictions and uncertainty by learning based on a limited number of current knowledge states (i.e., particles), it *never* reconsiders past categorization decisions. In many linguistic tasks, however, learners revisit past decisions (Frazier and Rayner, 1982) and intuitively we would expect categories to change based on novel evidence, especially in the early learning phase. In fixed-lag smoothing, a particle smoothing variant, model updates include systematic revision of a fixed set of previous observations in the light of newly encountered evidence (Briers et al., 2010). Based on this framework, we will investigate different schemes for informed sequential learning.

Finally, we would like to compare the model's predictions against behavioral data, and examine more thoroughly how categories and features evolve over time.

Acknowledgments We would like to thank Charles Sutton and members of the ILCC at the School of Informatics for their valuable feedback. We acknowledge the support of EPSRC through project grant EP/I037415/1.

References

Agirre, Eneko and Aitor Soroa. 2007. Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the 4th International Workshop on Semantic Evaluations*. Prague, Czech Republic, pages 7–12.

Anderson, John R. 1991. The adaptive nature of human categorization. *Psychological Review* 98:409–429.

Biemann, Chris. 2006. Chinese Whispers - an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of TextGraphs: the 1st Workshop on Graph Based Methods for Natural Language Processing*. New York City, pages 73–80.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Bomba, Paul C. and Eimas R. Siqueland. 1983. The nature and structure of infant form categories. *Journal of Experimental Child Psychology* 35:294–328.

Borschinger, Benjamin and Mark Johnson. 2011. A particle filter algorithm for Bayesian word segmentation. In *Proceedings of the Australasian Language Technology Association Workshop*. Canberra, Australia, pages 10–18.

Briers, Mark, Arnaud Doucet, and Simon Maskell. 2010. Smoothing algorithms for state-space models. *Annals of the Institute of Statistical Mathematics* 62(1):61–89.

Brody, Samuel and Mirella Lapata. 2009. Bayesian word sense induction. In *Proceedings of the 12th Conference of the European Chapter of the ACL*. Athens, Greece, pages 103–111.

Cree, George S., Ken McRae, and Chris McNorgan. 1999. An attractor model of lexical conceptual processing: Simulating semantic priming. *Cognitive Science* 23(3):371–414.

Douc, Randal. 2005. Comparison of resampling schemes for particle filtering. In *4th International Symposium on Image and Signal Processing and Analysis*. Zagreb, Croatia, pages 64–69.

Doucet, Arnaud, Nando de Freitas, and Neil Gordon. 2001. *Sequential Monte Carlo Methods in Practice*. Springer, New York.

Fearnhead, Paul. 2004. Particle filters for mixture models with an unknown number of components. *Statistics and Computing* 14(1):11–21.

Fountain, Trevor and Mirella Lapata. 2010. Meaning representation in natural language categorization. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*. Portland, Oregon, pages 1916–1921.

Fountain, Trevor and Mirella Lapata. 2011. Incremental models of natural language category acquisition. In *Proceedings of the 33nd Annual Conference of the Cognitive Science Society*. Boston, Massachusetts, pages 255–260.

Frazier, Lyn and Keith Rayner. 1982. Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology* 14(2):178–210.

Geman, Stuart and Donald Geman. 1984. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6(6):721–741.

Griffiths, Thomas L., Kevin R. Canini, Adam N. Sanborn, and Daniel J. Navarro. 2007. Unifying rational models of categorization via the hierarchical Dirichlet process. In *Proceedings of the 29th Annual Conference of the Cognitive Science Society*. Nashville, Tennessee, pages 323–328.

Griffiths, Thomas L., Adam N. Sanborn, Kevin R. Canini, John D. Navarro, and Joshua B. Tenenbaum. 2011. Nonparametric Bayesian models of categorization. In Emmanuel M. Pothos and Andy J. Wills, editors, *Formal Approaches in Categorization*, Cambridge University Press, pages 173–198.

Harris, Zellig. 1954. Distributional structure. *Word* 10(23):146–162.

Jern, Alan and Charles Kemp. 2013. A probabilistic account of exemplar and category generation. *Cognitive Psychology* 66:85–125.

Kemp, Charles, Patrick Shafto, and Joshua B. Tenenbaum. 2012. An integrated account of generalization across objects and features. *Cognitive Psychology* 64:35–75.

Lang, Joel and Mirella Lapata. 2011. Unsupervised semantic role induction with graph partitioning. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK., pages 1320–1331.

Levy, Roger P., Florencia Reali, and Thomas L. Griffiths. 2009. Modeling the effects of mem-

ory on human online sentence processing with particle filters. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 937–944.

McCallum, Andrew Kachites. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

McRae, Ken, George S. Cree, Mark S. Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavioral Research Methods* 37(4):547–59.

Medin, Douglas L. and Marguerite M. Schaffer. 1978. Context theory of classification learning. *Psychological Review* 85(3):207–238.

Nosofsky, Robert M. 1988. Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14:700–708.

Posner, Michael I. and Steven W. Keele. 1968. On the genesis of abstract ideas. *Journal of Experimental Psychology* 21:367–379.

Quinn, Paul C. and Peter D. Eimas. 1996. Perceptual cues that permit categorical differentiation of animal species by infants. *Journal of Experimental Child Psychology* 63:189–211.

Reed, Stephen K. 1972. Pattern recognition and categorization. *Cognitive psychology* 3(3):382–407.

Riordan, Brian and Michael N. Jones. 2011. Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science* 3(2):303–345.

Rosenberg, Andrew and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Prague, Czech Republic, pages 410–420.

Sanborn, Adam N., Thomas L. Griffiths, and Daniel J. Navarro. 2006. A more rational model of categorization. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*. Vancouver, Canada, pages 726–731.

Smith, Edward E. and Douglas L. Medin. 1981. *Categories and Concepts*. Harvard University Press, Cambridge, MA, USA.

Starkey, David. 1981. The origins of concept formation: Object sorting and object preference in early infancy. *Child Development* pages 489–497.

Storms, Gert, Paul De Boeck, and Wim Ruts. 2000. Prototype and exemplar-based information in natural language categories. *Journal of Memory and Language* 42:51–73.

Vinson, David and Gabriella Vigliocco. 2008. Semantic feature production norms for a large set of objects and events. *Behavior Research Methods* 40(1):183–190.

Voorspoels, Wouter, Wolf Vanpaemel, and Gert Storms. 2008. Exemplars and prototypes in natural language concepts: A typicality-based evaluation. *Psychonomic Bulletin & Review* 15(3):630–637.

Yao, Xuchen and Benjamin Van Durme. 2011. Nonparametric Bayesian word sense induction. In *Proceedings of TextGraphs-6: Graph-based Methods for Natural Language Processing*. Portland, Oregon, pages 10–14.

Zeigenfuse, Matthew D. and Michael D. Lee. 2010. Finding the features that represent stimuli. *Acta Psychologica* 133(3):283–295.